

## BAB II

### LANDASAN TEORI

#### 2.1. Text Mining

*Text Mining* merupakan suatu cara untuk menemukan informasi atau cara baru yang belum ada sebelumnya dengan melakukan proses dan analisa pada data berjumlah besar. Dalam melakukan analisa sebagian atau keseluruhan *unstructured text*, *text mining* dapat mengasosiasikan teks dengan yang lainnya didasari dengan aturan – aturan tertentu[10]. *Text Mining* juga dapat didefinisikan sebagai “*penemuan informasi baru dan tidak diketahui sebelumnya oleh computer, dengan secara otomatis mengekstrak informasi dari sumber – sumber teks tak terstruktur yang berbeda*”[11], [12]. Sumber data pada *Text Mining* berasal dari tulisan atau teks yang relatif tidak terstruktur karena menerapkan tata bahasa manusia atau *natural language*[12].

*Text mining* juga terdapat beberapa tipe menurut Dean Abbot[13] yaitu:

1. *Search and Information Retrieval (IR)*

Menyimpan dan mengambil data text, meliputi search engines atau mesin pencari dan kata kunci.

2. *Document Clustering*

Menggabungkan dan mengkategorikan istilah, potongan paragraf atau dokumen menggunakan metode *clustering* data mining.

3. *Document Classification*

Penggabungan dan pengkategorian potongan paragraf atau dokumen menggunakan metode klasifikasi berdasarkan pada trained model pada contoh label.

#### 4. *Web Mining*

Data dan text mining pada internet dengan spesifikasi khusus yang berfokus pada skala dan keterkaitan terhadap suatu web.

#### 5. *Information Extraction*

Identifikasi dan mengekstrak fakta yang relevan dan hubungan dari data text yang tidak terstruktur; proses pembuatan data terstruktur dari data tidak terstruktur dan semi-terstruktur.

#### 6. *Natural Language Processing*

Bahasa prosesing tingkat rendah dan pemahaman tugas (seperti, pengambilan bagian pidato); yang biasa digunakan secara sinonim dengan komputasi linguistik.

#### 7. *Concept Extraction*

Penggabungan kata – kata dan frasa kedalam kelompok yang mirip.

### **2.2. Analisis Sentimen**

Analisa sentimen atau *opinion mining*, merupakan sebuah *Natural Language Processing, Text Analytics* untuk melakukan identifikasi dan melakukan klasifikasi opini pada beberapa sumber data. Analisa sentimen juga berfokus pada opini seorang penulis terhadap sebuah topik. Pada dasarnya, klasifikasi analisa sentimen

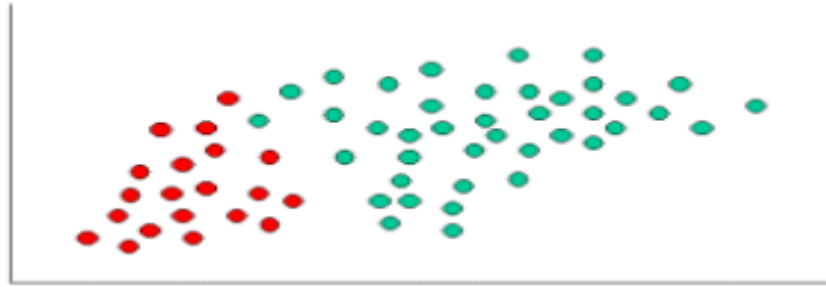
menggunakan ekspresi dari sebuah text pada sumber tertentu dengan dua tipe yaitu, fakta objektif tentang sebuah peristiwa dan yang kedua adalah opini yang merupakan sebuah ekspresi sentimen, attitude, dan ungkapan sebuah emosi[14]. Salah satu contoh penggunaan atau pemanfaatan analisa sentimen adalah melakukan identifikasi kecenderungan terhadap pasar dan opini pasar suatu barang.

*Natural Language Processing* merupakan sebuah metode yang dapat digunakan untuk menerapkan option mining. *Part-Of-Speech Tagging* yang biasa disebut dengan *POS Tagging* dapat digunakan untuk memberikan kelas kata (tag) terhadap setiap kata suatu teks. Penelitian yang ditujukan untuk mengembangkan sistem *POS Tagging* yang dilakukan menggunakan *Hidden Markov Model*. Akurasi yang didapatkan bisa mencapai angka 85% hingga angka 96%[14].

*Opinion Mining* atau analisa sentimen biasa merupakan kombinasi *text mining* dengan bahasa yang sering digunakan atau *natural language processing*. Metode *text mining* yang biasa digunakan untuk menyelesaikan masalah *opinion mining* adalah *Naïve Bayes Classifier* dan *K-Nearest Neighbor* (KNN). *Naïve Bayes Classifier* dan KNN dapat digunakan untuk melakukan klasifikasi opini kedalam beberapa kelompok. Kedua metode tersebut juga dapat mengklasifikasikan opini positif dan negatif. Dan juga, kedua metode tersebut dapat berfungsi dengan baik dan maksimal untuk metode pengklasifikasian teks.

### 2.3. Naïve Bayes Classifier

*Naïve Bayes Classifier* merupakan salah satu metode probabilitas sederhana yang melakukan perhitungan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan[15]. *Naïve Bayes Classifier* juga memiliki cara pengklasifikasian yaitu memiliki asumsi bahwa efek dari suatu nilai atribut tertentu tidak tergantung (independen) terhadap nilai atribut lainnya[16], [17].



**Gambar 2. 1. Naïve Bayes Classifier**

Pengklasifikasian menggunakan metode *Naïve Bayes Classifier* dapat sangat efisien dan akurat, terutama ketika jumlah variabel terbilang tinggi. Contoh sederhana perhitungan *Naïve Bayes Classifier* misalnya terdapat pada gambar 2.1, pada gambar terdapat dua kumpulan data hijau dan merah [17]. Data yang baru ditambahkan akan ditentukan merupakan bagian data hijau atau data merah. Karena data hijau lebih banyak dari data merah, maka akan diasumsikan data yang baru memiliki probabilitas menjadi anggota hijau dua kali lebih besar dari merah.

$$\text{Probabilitas untuk titik berwarna hijau}[17] = \frac{\text{Jumlah data hijau}}{\text{Jumlah data Keseluruhan}}$$

$$\text{Probabilitas untuk titik berwarna merah[17]} = \frac{\text{Jumlah data merah}}{\text{Jumlah data keseluruhan}}$$

*Naïve Bayes Classifier* mengasumsikan penyederhanaan bahwa nilai atribut saling bebas jika diberikan nilai outputnya. Probabilitas dapat diamati secara bersama dengan produk dari probabilitas individu. *Naïve Bayes Classifier* sering kali bekerja lebih baik dalam banyak situasi pada dunia nyata yang tidak tentu[15], [18].

Persamaan pada teorema *Naïve Bayes*:

$$P(H|X) = \frac{P(X|H).P(H)}{P(H)} \dots$$

#### **Rumus 2. 1. Persamaan Naïve bayes**

Penjelasan Rumus Persamaan *Naïve Bayes*[15]:

X = Data dengan Class yang belum diketahui.

H = Hipotesis data menggunakan suatu Class yang spesifik.

$P(H|X)$  = Probabilitas hipotesis H berdasarkan kondisi X (*Parteriori Probabilitas*).

$P(H)$  = Probabilitas hipotesis H (*Prior Probabilitas*).

$P(X|H)$  = Probabilitas X berdasarkan kondisi pada hipotesis H.

$P(X)$  = Probabilitas H

Untuk memproses sebuah klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas yang cocok bagi hasil analisa yang diinginkan. Karena itu, metode diatas disesuaikan sebagai berikut[15], [18]:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)}$$

**Rumus 2. 2. Contoh perhitungan Naïve Bayes**

Variabel C merupakan kelas, variabel F1...Fn merupakan karakteristik petunjuk yang dibutuhkan untuk menentukan klasifikasi. Maka, rumus diatas menyebutkan bahwa peluang masuknya sebuah sampel karakteristik dalam kelas C (*Posterior*) adalah peluang munculnya kelas C (*Prior*), dikalikan dengan peluang kemunculan sebuah karakteristik sampel pada kelas C (*likelihood*), kemudian dibagi dengan peluang munculnya karakteristik secara global (*evidence*). Rumus diatas dapat ditulis secara sederhana seperti berikut [17], [18]:

$$posterior = \frac{prior \times likelihood}{evidence}$$

**Rumus 2. 3. Perhitungan posterior**

Nilai *evidence* selalu tetap untuk setiap kelas pada suatu sampel. Nilai yang diperoleh dari *posterior* nantinya akan dibandingkan dengan suatu nilai *posterior* lainnya untuk menentukan masuknya suatu sampel kepada kelas saat melakukan klasifikasi[15].

## 2.4. *K-Nearest Neighbor*

*K-Nearest Neighbor* yang biasa disingkat menjadi *KNN* merupakan sebuah metode algoritma *supervised*. *Supervised Learning* merupakan sebuah proses untuk menemukan sebuah pola baru pada sebuah data dengan menghubungkan pola data yang ada dengan data yang belum ada atau data baru. Sedangkan, *Unsupervised Learning* bertujuan untuk, menemukan pola dalam sebuah data. Tujuan utama dari algoritma *KNN* adalah berfokus untuk klasifikasi sebuah objek berdasarkan atribut dan *training samples* atau biasa disebut dengan data *training*. Hasil pengklasifikasian algoritma *KNN* yang diklasifikasikan berdasarkan atribut mayoritas pada kategori *KNN*. Hal yang paling penting dalam algoritma *KNN* adalah algoritma ini menggunakan tetangga terdekat sebagai nilai dari sampel yang diuji[19].

Pada algoritma *KNN* terdapat perhitungan jarak yang menggunakan *Euclidian Distance*. Jarak ini merupakan jarak umum yang paling banyak digunakan pada data numerik. *Euclidian Distance* dapat diartikan sebagai berikut[19]:

$$d(x_i, x_j) = \sqrt{\sum_r^n (ar(x_i) - ar(x_j))^2}$$

### Rumus 2. 4. Euclidean Distance

Keterangan terkait Rumus *Euclidian Distance*:

$d(x_i, x_j)$  = Jarak *Euclidian Distance*

$(x_i)$  = *record* ke – i

$(x_j)$  = record ke - j

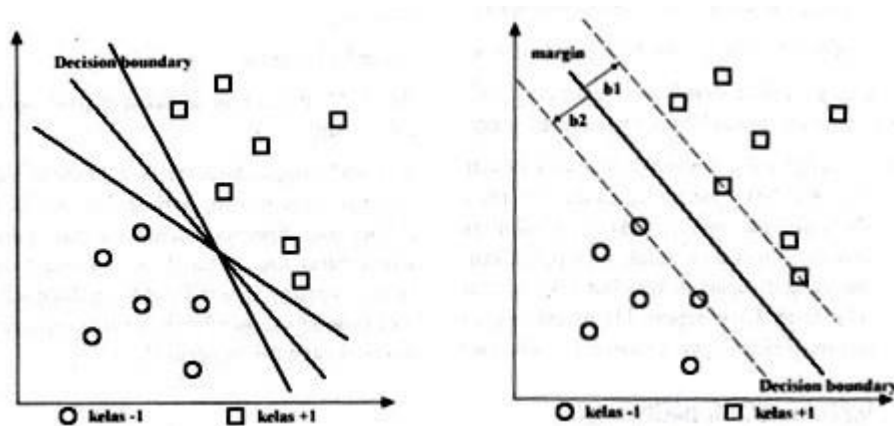
$(a_r)$  = data ke - r

$ij$  = 1, 2, 3, ... n

*KNN* adalah sebuah algoritma yang berfokus pada penentuan jarak saat melakukan pengujian *data testing* dengan *data training* berdasarkan nilai paling kecil dari tetangga terdekat[19].

## 2.5. Support Vector Machine

*Support Vector Machine* (SVM) merupakan salah satu metode klasifikasi. Konsep dasar pada metode *Support Vector Machine* adalah memaksimalkan batas *hyperplane* yang memisahkan suatu set data[8]. *Support Vector Machine* adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi linear dalam sebuah *feature space* (ruang fitur) berdimensi tinggi[20]. Berikut merupakan ilustrasi *Support Vector Machine*[8]:



Gambar 2. 2. Ilustrasi Metode Support Vector Machine



Pada gambar 2.2 terdapat konsep klasifikasi dengan menggunakan metode *Support Vector Machine*. Terlihat pada gambar pada sebelah kiri terdapat dua buah data yang di bagi menjadi kelas -1 dan kelas +1, terdapat juga *hyperplane* yang didefinisikan dengan tiga buah garis. Gambar sebelah kanan merupakan *hyperplane* yang paling maksimal karena hanya terdapat satu buah garis. Perhitungan *hyperplane* dapat dilakukan dengan menghitung jarak *margin* dengan data terdekat dari tiap kelas.[8]. *Support Vector Machine* juga memiliki beberapa karakteristik sebagai berikut[21]:

1. *Support Vector Machine* secara prinsip merupakan metode linear classifier.
2. *Pattern recognition* dilakukan dengan mentransformasikan data pada *input space* ke ruang yang berdimensi lebih tinggi, dan optimasi dilakukan pada ruang *vector* yang baru tersebut. Hal ini membedakan *Support Vector Machine* dari solusi *pattern recognition* pada umumnya, pada ruang hasil transformasi yang berdimensi lebih rendah daripada dimensi *input space*.
3. Menerapkan strategi *Structural Risk Minimization* (SRM).
4. Prinsip kerja *Support Vector Machine* pada dasarnya hanya mampu menangani klasifikasi dua kelas.

## **2.6. Confusion Matrix**

*Confusion matrix* merupakan pengukuran performa pada klasifikasi *machine learning*, yang merupakan tabel dengan 4 kombinasi dari nilai prediksi dan nilai *actual*.

Nilai prediksi merupakan keluaran dari program yang nilainya berupa positif dan negatif, sedangkan nilai *actual* adalah nilai yang sebenarnya berupa *true* dan *false*[22].

Berikut merupakan contoh confusion matrix untuk dua kelas klasifikasi[23]:

**Tabel 2. 1. Tabel Confusion Matrix**

		<i>Predicted</i>	
		<i>Positive “+”</i>	<i>Negative “-”</i>
<i>Actual</i>	<i>Positive (+)</i>	TP	FN
	<i>Negative (-)</i>	FP	TN

Tabel 2.1 diatas dapat diartikan sebagai berikut[23]:

1. *True Positive* (TP) merupakan prediksi positif dan benar.
2. *True Negative* (TN) merupakan prediksi negatif dan benar.
3. *False Positive* (FP) merupakan prediksi positif dan salah.
4. *False Negative* (FN) merupakan prediksi negatif dan salah.

Perhitungan akurasi, presisi, dan *recall* yang dilakukan dengan *confusion matrix* sebagai berikut[24] pada tabel 2.2:

**Tabel 2. 2. Tabel Persamaan Confusion Matrix**

Akurasi	$(TP + TN) / (TP + FP + FN + TN)$
Presisi	$TP / (TP + FP)$
Recall	$TP / (TP + FN)$

Akurasi menggambarkan seberapa kuat suatu model dalam melakukan klasifikasi dengan benar, presisi menggambarkan akurasi antara data dengan hasil prediksi yang dilakukan oleh model klasifikasi, dan *recall* menggambarkan keberhasilan model dalam memprediksi atau menemukan sebuah informasi[22].

Confusion matrix memiliki beberapa manfaat yaitu[25]:

1. Menunjukkan bagaimana model ketika membuat prediksi.
2. Tidak hanya memberi informasi tentang kesalahan yang dibuat oleh model tetapi juga jenis kesalahan yang dibuat.
3. Setiap kolom dari *confusion matrix* mempresentasikan *instance* dari kelas prediksi.
4. Setiap baris dari *confusion matrix* mewakili *instance* dari kelas aktual.

## **2.7. Media Sosial**

Media Sosial menurut Andreas Kaplan dan Michael Haenlein merupakan sebuah kelompok aplikasi berbasis internet yang dibangun diatas dasar ideologi dan teknologi *Web 2.0*, dan memungkinkan penciptaan dan pertukaran *user-generated content*[26]. *Web 2.0* menjadi *platform* dasar media sosial. Menurut Kaplan dan Haelein ada enam jenis media sosial: proyek kolaborasi (misalnya, Wikipedia), *blog* dan *microblogs* (misalnya, *Twitter*), komunitas konten (misalnya, youtube), situs jaringan sosial (misalnya, facebook dan Instagram), virtual game (misalnya, world of warcraft), dan virtual sosial (misalnya, second life)[26].

Media sosial adalah sebuah media online, dengan para penggunanya bisa dengan mudah berpartisipasi, berbagi dan menciptakan isi meliputi blog, jejaring sosial, wiki, forum, dan dunia virtual. Blog, jejaring sosial, dan Wiki merupakan bentuk media sosial yang paling umum digunakan oleh masyarakat di seluruh dunia.[26]

Media sosial juga memiliki beberapa fungsi [26] yaitu:

1. Sosial media adalah media yang didesain untuk memperluas interaksi sosial manusia menggunakan internet dan teknologi *web*.
2. Sosial media berhasil mentransformasikan praktik komunikasi searah media siaran dari satu institusi media ke banyak audience menjadi praktik komunikasi dialog antar banyak *audience*.
3. Sosial media mendukung demokratisasi pengetahuan dan informasi. Mentransformasikan manusia dari pengguna isi pesan menjadi pembuat pesan itu sendiri.

## **2.8. Twitter**

*Twitter* merupakan sebuah situs jejaring sosial yang berdiri pada 21 Maret 2006 oleh Jack Dorsey, Noah Glass, Biz Stone, dan Evan Williams. *Twitter* sendiri dapat digunakan dan dinikmati oleh publik pada 15 Juli 2006 yang bermarkas di San Francisco, California, Amerika Serikat[27].

*Twitter* adalah sebuah layanan jejaring sosial atau lebih dikenal sebagai media sosial pada saat ini yang dapat digunakan oleh pengguna sebagai sarana berkirim pesan

yang isinya tidak lebih dari 140 karakter. Namun, pada 7 November 2017 *Twitter* menambah kapasitas pesan menjadi 280 karakter yang mulai populer sejak tahun 2007[27].

Beberapa fitur – fitur penting pada *Twitter* antara lain:

1. *Home*, merupakan halaman pertama yang muncul pada *Twitter* setelah kita melakukan *Log-in*. Halaman ini berisi kumpulan *Tweet* yang di sebarakan oleh pengguna seperti berita, informasi, gambar, suara, bahkan video.
2. *Profile*, merupakan informasi pengguna mengenai data diri, asal negara, dan merupakan sebuah kumpulan *Tweet* yang pernah disebarakan oleh pengguna.
3. *Followers*, merupakan pengguna lain yang mengikuti kita pada *Twitter* yang disebut sebagai *Followers*.
4. *Following*, merupakan pengguna yang kita ikuti untuk melihat aktivitas yang dibagikan pada *Tweet* yang disebarakan oleh pengguna lainnya.
5. *Mentions*, merupakan sebuah fitur dalam *Twitter* yang melibatkan pengguna lainnya pada sebuah *Tweet*. Jika sebuah *Tweet* berisi nama akun dari pengguna lain, maka pengguna tersebut mendapatkan notifikasi.
6. *Direct Message*, merupakan fitur pesan private antar pengguna sehingga menjaga privasi terhadap pesan yang ingin dikirim.
7. *Search*, merupakan fitur ini merupakan fitur untuk mencari berbagai *Tweet* yang kita inginkan sesuai dengan topik yang ingin dilihat.

8. *Hashtag*, memiliki simbol depan “#” tanda pagar, yang berfungsi untuk mencari topik tertentu pada fitur *Search*.
9. *Trending Topic*, merupakan topik yang sedang ramai dibicarakan oleh banyak pengguna *Twitter* yang dibagikan melalui sebuah hashtag. Kita juga dapat mencari sebuah berita yang sedang hangat pada fitur *Search*.

### **2.9. Term Frequency Inverse Document Frequency (TF-IDF)**

Metode *TF-IDF* merupakan sebuah metode untuk menghitung bobot pada kata yang sering atau paling umum digunakan pada *information retrieval*. Metode ini adalah pemberian bobot hubungan pada kata (*term*) terhadap kata dalam dokumen, dan juga merupakan sebuah ukuran statistik yang dapat digunakan untuk melakukan evaluasi seberapa penting sebuah kata dalam dokumen. Frekuensi dalam kata pada dokumen diberikan untuk menunjukkan seberapa sering kata tersebut digunakan. Pada metode ini digunakan rumus untuk menghitung bobot yaitu[28]:

$$w_{dt} = t_{fdt} * i_{dft}$$

#### **Rumus 2. 5. Perhitungan bobot**

Penjelasan:

- $w_{dt}$  = merupakan bobot dokumen ke-d terhadap kata ke-t
- $t_{fdt}$  = merupakan banyaknya kata yang dicari dalam dokumen
- $i_{dft}$  = *Inversed Document Frequency* ( $\log (N/df)$ )
- $N$  = merupakan total dokumen
- $df$  = merupakan banyaknya dokumen yang mengandung kata yang dicari

### 2.10. Area Under Curve

*Area Under Curve* atau biasa disebut dengan *AUC* merupakan nilai yang berada pada *receiver operating characteristic* atau *ROC*, yang merupakan kurva yang diperoleh antara sensitivitas dan spesifitas pada suatu titik potong, *AUC* juga merupakan area luas yang berada dibawah kurva *ROC* atau merupakan keseluruhan dari fungsi *ROC*. Nilai pada *AUC* berada antara 0 dan 1. Nilai yang dihasilkan *AUC* dapat memberikan nilai pendukung tentang pengukuran yang dilakukan suatu model. Semakin besar nilai *AUC* yaitu menyentuh angka 1, maka semakin baik model yang digunakan dalam melakukan prediksi[29]. Nilai pada *AUC* adalah sebagai berikut:

1. Nilai 0,90 – 1,00 = *Excellent Classification*
2. Nilai 0,80 – 0,90 = *Good Classification*
3. Nilai 0,70 – 0,80 = *Fair Classification*
4. Nilai 0,60 – 0,70 = *Poor Classification*
5. Nilai 0,50 – 0,60 = *Failure Classification*

### 2.11. R Studio

*R Studio* merupakan sebuah program yang dikembangkan oleh R Core Team, yang dikenal dengan bahasa pemrograman *R*. Bahasa pemrograman ini merupakan salah satu *powerfull software* untuk melakukan analisa pada data. *R* diciptakan dengan tujuan awal untuk melakukan komputasi pada statistik dan grafik, yang digunakan oleh para ilmuwan untuk melakukan riset dan juga pada bidang akademis. Seiring perkembangan teknologi, kemampuan *R* sebagai bahasa pemrograman menjadi lebih

berkembang, contohnya adalah membuat aplikasi *website* interaktif atau *dashboard* dengan memanfaatkan *package* yang telah disediakan oleh *R Studio*[30]. Keuntungan penggunaan *R Studio* adalah telah tersedia banyak *package*, mudah dalam melakukan transformasi data, mampu menghasilkan grafik yang sangat bagus[30].

## 2.12. Rapid Miner

*RapidMiner* merupakan sebuah aplikasi perangkat lunak atau software yang bersifat open source (terbuka dan bebas untuk digunakan). *RapidMiner* merupakan sebuah aplikasi untuk melakukan analisis terhadap *data mining*, *text mining*, dan melakukan analisis prediksi. *RapidMiner* sebelumnya bernama YALE (Yet Another Learning Environment) yang dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Merisw, dan Simon Fischer pada University of Dortmund. *RapidMiner* merupakan aplikasi open source untuk melakukan *data mining* maupun *text mining* yang sudah digunakan lebih dari 40 negara[31].

*RapidMiner* sudah menyediakan *Graphic User Interface* (GUI) yang akan menghasilkan sebuah file *Extensible Markup Language* (XML) yang dapat diterapkan pada proses analisis.

*RapidMiner* juga memiliki beberapa manfaat yaitu[31]:

1. Berbahasa pemrograman Java sebagai sistem operasi (OS).
2. Proses pencarian dimodelkan sebagai *Trees*.
3. Representasi *XML* internal untuk memastikan format standar pertukaran data.



4. Multi-layer konsep yang menjadikan tampilan data lebih efisien.
5. Dilengkapi dengan *GUI*, *Command Line*, dan *JAVA API* yang dapat digabungkan dengan penggunaan program atau aplikasi lainnya.

RapidMiner juga memiliki beberapa fitur menarik yaitu[31]:

1. Memiliki berbagai macam algoritma *data mining*.
2. Memiliki bentuk grafik yang menarik seperti diagram histogram, *Tree Chart*, *3D Scatter Plots*.
3. Memiliki berbagai macam variasi *plugin*.
4. Telah menyediakan prosedur *data mining* dan *machine learning*.
5. *Extraction, Transformation, Loading* (ETL) telah tersedia pada *RapidMiner* serta data preprocessing, visualisasi, modelling dan evaluasi.
6. Dapat terintegrasi dengan program *data mining* seperti statistic *R*.

### **2.13. Penelitian Terdahulu**

Dari pencarian yang dilakukan melalui media Internet tentang penggunaan Sentiment Analisis yang menggunakan media sosial sebagai sarana penelitian, maka ditemukan beberapa jurnal dan penelitian sebagai berikut:

**Tabel 2. 3. Tabel Penelitian Terdahulu**

<b>Judul Penelitian, Peneliti, dan Tahun Publikasi</b>	<b>Objek Penelitian</b>	<b>Teknik Penelitian</b>	<b>Hasil Penelitian</b>
C. Destitus, W. Wella, and S. Suryasari, “Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia,” <i>Ultim. InfoSys J. Ilmu Sist. Inf.</i> , vol. 11, no. 2, pp. 107–111, 2020, doi: 10.31937/si.v11i2.1740.	Hasil akurasi metode <i>Support Vector Machine</i> dan <i>Information Gain</i> dari data <i>tweet</i> yang mengandung konten <i>cyberbullying</i>	<ul style="list-style-type: none"> <li>- Twitter</li> <li>- Informati on Gain</li> <li>- Support Vector Machine</li> <li>-</li> </ul>	<b>Metode Support Vector Machine mendapatkan hasil akurasi 80%, presisi 75,1%, dan recall 96%. Metode Information Gain mendapatkan akurasi 86%, presisi 81%, recall 96%. Sedangkan nilai <i>f-measure</i> masing – masing sebesar 85% dan 87%.</b>
Nomleni, P., Hariadi, M., & Purnama, I. K. E. (2014). Sentiment Analysis Berbasis Big Data. Seminar Nasional Rekayasa Teknologi Industri Dan Informasi, 9, 142–149.	Keluhan masyarakat terhadap pemerintah pada media sosial Facebook dan Twitter sapawarga data berbahasa Indonesia.	<ul style="list-style-type: none"> <li>- Sentiment Analysis</li> <li>- Hadoop</li> <li>- Support Vector Machine (SVM)</li> </ul>	<b>Analisa dilakukan menggunakan <i>Support Vector Machine</i> dan memiliki tingkat akurasi klasifikasi yang semakin tinggi bila data yang digunakan semakin banyak</b>

<b>Judul Penelitian, Peneliti, dan Tahun Publikasi</b>	<b>Objek Penelitian</b>	<b>- Teknik Penelitian</b>	<b>- Hasil Penelitian</b>
Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. <i>Concurrency Computation</i> , 32(1), 1–29. <a href="https://doi.org/10.1002/cpe.5107">https://doi.org/10.1002/cpe.5107</a>	Pemahaman terhadap kelayakan, ruang lingkup dan relevansi aliansi menggunakan Soft Computing teknik Sentiment Analysis pada media sosial Twitter	<ul style="list-style-type: none"> <li>- Sentiment Analysis</li> <li>- Soft Computing</li> <li>- Twitter</li> </ul>	<ul style="list-style-type: none"> <li>- Mining, Analyzing dan classifying sentimen dari sebuah web lebih sulit dilakukan secara komputasi.</li> <li>- Analisa sentimen menggunakan pendekatan algoritma <i>Fuzzy Logic</i> seperti model tipe 2 adalah sebuah dimensi baru yang terbuka untuk eksplorasi lebih lanjut.</li> </ul>
Palupi, E. S., Pahlevi, S. M., Bina, U., Informatika, S., Magister, P., & Komputer, I. (2020). Inti nusa mandiri. <i>Inti Nusa Mandiri</i> , 14(2), 133–138. <a href="https://doi.org/https://doi.org/10.33480/inti.v14i2.1178">https://doi.org/https://doi.org/10.33480/inti.v14i2.1178</a> VOL.	Sentiment Analysis opini publik mengenai <i>COVID-19</i> pada Twitter	<ul style="list-style-type: none"> <li>- Twitter</li> <li>- <i>Naïve Bayes</i></li> <li>- <i>K-Nearest Neighbor</i></li> </ul>	<ul style="list-style-type: none"> <li>- <i>COVID-19</i> merupakan penyakit yang tidak sedikit meresahkan masyarakat.</li> <li>- Hasil pengujian dilakukan menggunakan metode <i>Naïve Bayes</i> dan KNN dan mendapatkan hasil opini yang cenderung positif.</li> <li>- <i>Naïve Bayes</i> merupakan model klasifikasi yang paling akurat dibanding KNN dengan hasil akurasi 63.21%.</li> </ul>

<b>Judul Penelitian, Peneliti, dan Tahun Publikasi</b>	<b>Objek Penelitian</b>	<b>Teknik Penelitian</b>	<b>Hasil Penelitian</b>
KURNIAWAN, R., & APRILIANI, A. (2020). Analisis Sentimen Masyarakat Terhadap Virus Corona Berdasarkan Opini Dari Twitter Berbasis Web Scraper. In Jurnal INSTEK (Informatika Sains dan Teknologi) (Vol. 5, Issue 1, p. 67). <a href="https://doi.org/10.24242/instek.v5i1.13681">https://doi.org/10.24242/instek.v5i1.13681</a>	Bagaimana dampak virus corona di Indonesia sesuai opini masyarakat melalui Twitter	<ul style="list-style-type: none"> <li>- Text Mining</li> <li>- Analisis Sentimen</li> <li>- Twitter</li> <li>- Web Scraper</li> </ul>	<ul style="list-style-type: none"> <li>- Hasil pengamatan sentimen melalui <i>Twitter</i> mendapati opini masyarakat cenderung negatif terhadap topik corona yaitu mencapai 79%.</li> <li>- Polaritas tweet netral 11% dan positif hanya 10%.</li> <li>- Dapat disimpulkan bahwa corona memberikan dampak negatif berdasarkan hasil sentimen negatif yang lebih tinggi.</li> </ul>
Rossi, A., Lestari, T., Setya Perdana, R., & Fauzi, M. A. (2017). Analisis Sentimen Tentang Opini Pilkada DKI 2017 pada Dokumen Twitter Berbahasa Indonesia Menggunakan N��ive Bayes dan Pembobotan Emoji. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 1(12), 1718–1724. <a href="http://j-ptiik.ub.ac.id">http://j-ptiik.ub.ac.id</a>	Analisis Sentimen tentang opini pilkada DKI 2017 pada Twitter berbahasa Indonesia	<ul style="list-style-type: none"> <li>- Analisis Sentimen</li> <li>- Twitter</li> <li>- <i>Na��ve Bayes</i></li> </ul>	Hasil akurasi sistem pada pembobotan tekstual 68,52% dan non-tekstual 75,93% dan penggabungan keduanya sebesar 74,81%. Sehingga disimpulkan penggabungan kedua pembobotan dapat meningkatkan hasil akurasi sistem.

Pada tabel 2.6 tentang penelitian yang telah dilakukan terdahulu terdapat beberapa penelitian yang merupakan acuan pada pembuatan laporan ini. Penelitian

dengan judul “*Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia*” mendapatkan hasil penelitian dengan *Support Vector Machine* memiliki akurasi sebesar 80%, presisi 75,1%, dan recall 96%, sedangkan dengan *Information Gain* akurasi sebesar 86%, presisi 95%, dan recall 95% serta nilai *f-measure* sebesar 85% dan 87%. Pada penelitian yang berjudul “Keluhan masyarakat terhadap pemerintah pada media sosial *Facebook* dan *Twitter* sapawarga data berbahasa Indonesia” yang ditulis oleh [32] melakukan sebuah analisa sentimen menggunakan *Twitter* dan *Facebook* sebagai sumber pengumpulan data dengan memanfaatkan metode *Support Vector Machine* yang menghasilkan tingkat nilai akurasi yang tinggi jika data yang digunakan semakin banyak. Kemudian, penelitian kedua yang ditulis oleh [33], dengan judul penelitian “Sentiment Analysis opini publik mengenai *COVID-19* pada *Twitter*” yang menggunakan *Twitter* sebagai sumber data dan penggunaan algoritma *Naïve Bayes* dan *K-Nearest Neighbor* yang menunjukkan bahwa metode *Naïve Bayes* memiliki tingkat akurasi yang lebih tinggi yaitu sebesar 63.